## I. Overview

The National Oceanic and Atmospheric Administration (NOAA) conducts research and gathers data about the global oceans, atmosphere, space, and sun, and applies this knowledge to science and services that touch the lives of all Americans.

NOAA warns of dangerous weather, charts our seas and skies, guides our use and protection of ocean and coastal resources, and conducts research to improve our understanding and stewardship of the environment which sustains us all. NOAA's mission is embodied in its four strategic goals:

- Protect, restore, and manage the use of coastal and ocean resources through ecosystem-based management.
- Understand climate variability and change to enhance society's ability to plan and respond.
- Serve society's needs for weather and water information.
- Support the Nation's commerce with information for safe, efficient, and environmentally sound transportation.

To support the NOAA mission, Forecast Systems Laboratory (FSL) set up a data center in Boulder, Colorado to develop, test and evaluate new, advanced weather forecasting techniques and technologies, and transfer these to operational services. This data center provides the infrastructure for acquiring, processing, storing, and distributing large volumes of conventional (operational) and advanced meteorological data and products.

A major task, enabled by the data center located in Boulder, involves the development and running in real time of a variety of advanced atmospheric models. An important aspect of the modeling work is the analysis of large volumes of data and running the prognoses which require very substantial amounts of computational power. To meet these computational needs NOAA acquired a High Performance Computing System (HPCS) in 1999. This system is used to support projects throughout NOAA.

As the users of the HPCS continue to increase the spatial and temporal resolution of the applications it runs, a procurement of a replacement system for the HPCS is necessary. The replacement system will be known as the NOAA HPCS for Research Applications (NHRA) and will be used to run existing and newly-developed applications by the existing users as well as new users from within the NOAA research community. The NHRA will be housed at FSL as the HPCS currently is. NOAA seeks a balanced system including a Mass Storage System, network attached disks, communications interfaces, test platforms (if

necessary), and the computational platform. NOAA requires, at a minimum, an initial system and a final system. The initial system must meet or exceed the performance criteria cited in section IV of this Statement of Need. The final system shall provide a significant upgrade over the initial system and shall be in place no later than 36 months into the contract. There shall be no more than one intermediate upgrade between the initial and final systems.

II. Existing Infrastructure

Since the NHRA is intended to be a NOAA-wide resource, this document will only discuss the existing infrastructure relevant to the support of the NHRA within FSL's central facility. Currently, the HPCS consumes a majority of the floor-space, power, and cooling capacity available within FSL's primary computer room (2B201 in the David Skaggs Research Center located in Boulder, Colorado). The HPCS is composed of the following:

- 770 dual-processor 2.2GHz Intel Pentium IV-Xeons with 1GByte of memory each. These are interconnected with Myrinet2000 from MyriCom (fiber connections).
- 142 dual-processor 833MHz Alpha EV68s with 1 or 2GBytes of memory each. These are interconnected with Myrinet2000 from MyriCom (copper connections).
- 3 Compaq (HP) EV40 with 667MHz Alpha EV68s.
- 2 DataDirect Networks Silicon Storage Devices. These devices are FibreChannel attached.
- Approximately 12TBytes of Dell FibreChannel storage.
- 3 Brocade 16 port (1 gigabit) switches.
- A Mass Storage System (MSS) composed of a Sun E450, an SGI Origin 2000, an ADIC AML/J robot with 8 IBM LTO drives and 4 Sony AIT-3 drives. The disk cache is located on the larger of the two DataDirect Silicon Storage Devices. The total usage of disk space for the MSS on the DataDirect devices is 1.4 TBytes, over 1 TByte is cache the remainder is database storage. The software running on the Sun is ADIC's StorNext while the software running on the SGI is ADIC's FileServ and VolServ.



Currently, FSL's network infrastructure is Gigabit Ethernet based, scaled appropriately from Fast Ethernet (100 Mbps) to multiple Gigabit Ethernet trunks as needed. The laboratory is migrating all server-class systems to Gigabit Ethernet. It is also positioned to implement 10 Gigabit Ethernet when additional bandwidth becomes necessary for the future. FSL has a number of connections to public and research networks. It's primary high-speed connection is currently

via the Abilene network. FSL will continue to pursue high-speed connections such as National Lambda Rail, and high-performance technologies such as national scale grid computing. The NHRA must be able to connect initially with multiple jumbo packet enabled Gigabit Ethernet links and have an upgrade path that will be compatible with FSL's upgrade path.

## III. Overview of the NHRA

The computational platform of the NHRA will have a varied job mix. A number of real-time meteorological models will be run at specific times on the computational platform requiring guaranteed resources. Model developers will make use of the machine regularly while improving the models. Regular runs of multiple models on perturbed saved data for data denial experiments will consume a significant portion of the computational platform. The data for these data denial experiments will reside on the Mass Storage System (MSS) and will be migrated to/from the network attached disk subsystem. Software developers will make less frequent use of the machine while improving infrastructure items and the data distribution software. A large number of jobs that use a single-processor or a small number of processors will be run on the computational platform. The number of batch jobs run daily on the current computational platform is in the thousands. The Government does not anticipate that this job load will decrease, but rather expects it to increase. The runtime for jobs ranges from a few minutes to over 24 hours.

Currently the batch system in use is Sun Grid Engine (SGE). FSL has developed scripts to submit jobs to ensure users are notified at job submission time if the job being submitted exceeds the resources allocated for the user's project. FSL has also developed a job scheduler that has the following functionality:

- Each project has a unique limit on the number of CPUs it can use at any point in time.
- Each project is a member of a class that also has a limit on the number of CPUs that can be used by the class.
- The scheduler can stop scheduling while jobs are still submitted to the queuing system
- An overflow resource is available so that idle processors can be utilized (with limits)
- A user may be associated with multiple projects

In addition to the software supplied by the the current HPCS vendor, FSL has developed software monitors in collaboration with the current vendor. These include status monitors that automatically notify support staff of detected problems and monitors that automatically take bad nodes off line. It is expected

that the NHRA will have equivalent or superior batch software as well as system monitoring software. It should be noted that the scheduler and other software is free open source and can be made available to vendors who express interest. There is, however, no support from the incumbent or the Government promised or implied.

The Government is also interested in having a subset of the computational platform to support visualization. This subset would have direct login capability (that is, a user should not have to log into another portion of the NHRA before being able to access the visualization platform(s)) from the FSL LAN and have visibility of all computational platform file systems. The software would include IDL, VIS5D, NCAR graphics, TCL/TK, and open source packages as the Government sees fit. The licenses for the commercial packages would be supplied by the Government. Users would use the visualization portion of the machine for both interactive visualizations and batch production of graphics files. Users should be able to log directly into the visualization portion, ask for resources through the batch system, and be able to regularly schedule processes on the visualization portion of the computational platform.

The success of the computational platform is directly related to the success of the users of the NHRA. As a result, the Government will track the success rate of a number of well defined procedures. A procedure comprises one or more batch jobs that a user submits for a single purpose. A procedure is successful if no job fails due to any type of system failure. System failures include, but are not limited to, disk failures, file system software failures, batch system failures, and, of course, hardware failures. One carefully monitored procedure for the initial year of the contract shall be a mirror of the Rapid Update Cycle (RUC) (or its Weather Research and Forecasting (WRF) model equivalent) that runs operationally at the NWS/National Centers for Environmental Prediction (NCEP). Note that the RUC is simply an example of a procedure that will be carefully monitored, the Government is not limited to having only one carefully monitored procedure. The RUC runs hourly at NCEP and will do so on the computational platform of the NHRA. The success rate shall be measured as the number of procedures that were successful divided by the number of possible procedures. A running average with a 30 day window shall be the success rate measure. The number of possible procedures is defined as the number of procedures that could be attempted minus the number of procedures that would have occurred during either a scheduled down-time or FSL central facility outage. Any procedures that cannot be attempted because of upstream data failures, bugs in the codes, or over subscription of the NHRA are also excluded. A procedure is successful if ALL of the resulting output scheduled for that procedure is produced in a satisfactory time-frame. (It should be noted that the current RUC involves approximately 20 separate batch jobs including a large number of

single-processor jobs.) Some scripts may automatically attempt retries within a small time window. Retries are only feasible if the batch system provides sufficient information so that a script can determine if a batch job is queued, running, or has exited the batch system as well as determine the exit status of a batch job. The average success rate for carefully monitored tasks utilizing the NHRA shall be greater than or equal to 99%. The average succcess rate is determined as the sum of the individual success rate for each carefully monitored task divided by the number of carefully monitored tasks. Prior to acceptance testing, the Government will identify the initial well tested set of monitored tasks. As the applications suites on the NHRA evolve, the Government can unilaterally change the suite of monitored tasks.

The computational platform shall be available 99% of the time. The availability measure will be a 30-day running average. The sampling rate for availability shall be as frequent as is feasible while not significantly impacting the system. Availability will be the minimum availability of all components. For example, if the home file system is down, the system will be deemed to be 0% available. If the batch system determines that a computational resource is not available, then that computational resource is deemed unavailable even if the component is fully functional. Planned outages are excluded from the availability calculations.

The MSS of the NHRA must be available 99% of the time as well. This includes network access, disk storage, and robotic storage.

The Contractor(s) shall provide a means of determining the status of the NHRA and provide notification to the Government's and its own support staff in order to resolve failures in a timely manner.

The Government requires that hardware and software maintenance support included in the contract cover the 5-year life of the system.

The Government requires that the Computational Platform and the MSS both are capable of conforming to all Government IT security requirements and policies. The requirements and policies can be found on the website: http://nhra.fsl.noaa.gov by following the Secuirty Policies link.

The Government requires on-site hardware and software support during normal Government business hours for the 5-year system life. The Government will supply two 150 square foot offices and approximately 100 square feet of lab space for use by the Contractor. None of this space will have furniture or lab benches. The space will have up to 4 telephones as well as adequate LAN connections.

The Government will provide staff including one systems administrator, one network engineer, a user support liaison, and a user support specialist for the compute platform. The Government will also provide a part-time system administrator for the MSS. A team lead will also be provided by the Government. Staffing levels may be adjusted based upon changing requirements and funding.

The Government requires on-site training on usage of the NHRA. This should include thorough familiarization of up to 30 scientists and programmers in applications programming and system usage. The Government will require annual off- or on-site training for 3 staff members in systems administration. Additional applications programming training is required if an upgrade significantly affects applications development.

IV. Requirements

The Government will supply the following codes as elements of the benchmark suite:



- Rapid Update Cycle (RUC)
- Weather Research and Forecasting (WRF) Model
- Mesoscale Model version 5 (MM5)
- Medium Range Forecast (MRF) Model
- Regional Ocean Modeling System (ROMS)

Each of these codes will have associated pre- and post-processing applications which will use one or more processors. The Contractor will provide a performance guarantee for the Computational Platform in terms of the number of batch suites that will be delivered per day. To evaluate this performance guarantee, a maximum of two file systems may be used. The first file system will be the home file system which will contain users' configuration files and no data. The second file system will be the data file system where the bulk of the I/O activities shall occur. Failure to meet the performance guarantee cannot be remedied by adjusting the source code for any application. The source code used to produce the benchmark results will be submitted with the proposal and only bug fixes shall be allowed thereafter.

***The actual batch configurations are not ready at this time, but there will be more than one configuration. One configuration will be provided for the initial year of operations, other configurations may be provided for subsequent time periods. Elements of the batch suites will have maximum run-times associated with them. For example, the RUC must run within 55 minutes including pre- and post-processing.***

The proposed Computational Platform of the NHRA must meet the minimum performance requirements in Table I below. It is required that the NHRA either meet the performance requirement in the fourth column or the performance requirement in the fifth column. Additional consideration will be based upon the results of the batch benchmarks mentioned above.

**Table I**

| ***Benchmark*** | ***Performance / Percent of Current System*** | ***Performance / Percent of Current System*** | ***Performance Requirement / Percent of NHRA*** | ***Performance Requirement / Percent of NHRA*** |
|---|---|---|---|---|
| RUC – 20KM (hybcst_sp only) | 2384 seconds on 36 CPUs (2.2 GHz Xeon) | 1476 seconds on 71 CPUs (2.2 GHz Xeon) | 1400 seconds on 1/44$^{th}$ of initial computational platform | 700 seconds on 1/22$^{nd}$ of initial computational platform |
| WRF-10KM CONUS | ***TBD*** | ***TBD*** | ***TBD*** | ***TBD*** |
| MM5 | ***TBD*** | ***TBD*** | ***TBD*** | ***TBD*** |
| MM5-Chem | ***TBD*** | ***TBD*** | ***TBD*** | ***TBD*** |
| MRF | ***TBD*** | ***TBD*** | ***TBD*** | ***TBD*** |
| ROMS | 2240 seconds on 64 CPUs (2.2GHz Xeon) | 2818 seconds on 32 CPUs (2.2GHz Xeon) | 2200 seconds on 1/48$^{th}$ of initial computational platform | 1100 seconds on 1/24$^{th}$ of initial computational platform |

The Government will make available some of the components of the existing HPCS. These will be supplied as Government Furnished Equipment (GFE) if the Contractor chooses to use these components. If the Contractor chooses to use the GFE, the Contractor must maintain the equipment during its operation. The Table II below indicates what GFE is available.

**Table II**

| ***NHRA Element*** | ***Item*** | ***Details*** |
|---|---|---|
| MSS | ADIC AML/J Robot | 8 LTO Tape Drives<br>4 AIT3 Tape Drives<br>1976 AIT Tapes<br>>1300 LTO Tapes |
| MSS | Sun E450 | |
| MSS | DDN SSA (Portion) | 1.4 TBytes of storage |

| *NHRA Element* | *Item* | *Details* |
|---|---|---|
| Computational Platform | Gigabit Ethernet Switch | Extreme Summit 7i (28 Copper and 4 Fiber Ports) |
| Computational Platform | Gigabit Ethernet Switch | Extreme Black Diamond (64 Copper Ports) |
| Computational Platform | MyriCom Myrinet2000 | 17 M3-E128 Chassis<br>17 M3-M Monitoring Cards<br>107 M3-SW16-8F Eight Switched Fiber<br>39 M3-SPINE-8F Eight Fiber Spine<br>88 M3-SW16-4DM Four Ribbon Connector Spine<br>768 NICs |
| Computational Platform | DDN SSA (1 + Portion) | 8.4 TBytes of storage<br>6 Tbytes of storage |



The NHRA will have a minimum of 30 TBytes of disk storage space available for users' data, this is defined as users' storage. This storage shall be visible from all portions of the Computational Platform and shall be exportable via NFS and (optionally) other mechanism to other servers within the Government supplied computational facility, including the MSS. This storage is exclusive of any storage necessary to keep systems related files such as kernels, libraries, utilities, and compilers. The storage required to keep systems related files is defined as systems storage. There should be enough systems storage to save up to four versions of the system software.

The Mass Storage System must allow for 1 Petabyte of near-line storage. The Contractor may elect to supply a completely new MSS or upgrade the existing ADIC StorNext system. If the Contractor elects to provide a new system, all of the existing data on the StorNext system must be migrated to the new system within 6 months of acceptance. There is approximately 175 TBytes of data currently stored on the existing system. This amount of storage is in addition to the required minimum of 1 Petabyte. It should be noted that the Government intends to make two copies of all data and thus 2 PBytes of media is required. The storage capacity shall be determined without considering the compression capability of the hardware and software. The Government will consider

compression capability in its evaluation, but the minimum storage capacity must be met without compression.

The MSS and the Computational Platform must be proposed separately and the Government reserves the right to make two separate awards. A Contractor may propose either or both. If the Contractor proposes both an MSS and a Computational Platform, each must be able to operate independently of the other. The Contractor may indicate cost, performance, or capacity advantages if both contracts are awarded to the same Contractor.

The software functionality for the Computational Platform of the NHRA is described in Table III below. Failure to offer any **required** element will result in a deficient proposal. Additional consideration of a proposal will be given if **desired** items are offered. A lesser amount of additional consideration will be given for a **useful** item that is offered:

**Table III**

| ***Software Function*** | ***Feature*** | ***Importance*** |
|---|---|---|
| System Administration | Maintain consistent system images | Required |
| | Automated software upgrades | Required |
| Resource Partitioning | Isolate resources for user groups | Required |
| | Separate development from operations | Required |
| | Run different OS revisions in different partitions | Desired |
| Job Scheduling | Batch System | Required |
| | Grid-enabled Batch System | Desired |
| | Avoid resource contention between jobs | Required |
| Job Prioritization | High-priority jobs run first | Required |
| | Resource reservation | Required |
| | Low-priority jobs automatically suspend/swap-out if higher-priority job starts | Desired |
| Checkpoint-Restart | User-initiated via API | Desired |
| | Support MPI messages-in-flight | Desired |
| | Automatic or operator initiated | Useful |
| Fault Tolerance | Single-CPU failure does not require reboot of entire computational platform | Required |
| | Single-CPU failure only kills jobs running on failed CPU | Required |
| | Jobs can be migrated to other CPUs | Desired |

| *Software Function* | *Feature* | *Importance* |
|---|---|---|
| | Jobs automatically migrate to other CPUs if failure likely | Useful |

Table IV below describes minimum applications software components and features. Features are divided into two requirement categories: required and desired. Required features must be provided in the proposed initial system. Desired features are strongly desired by FSL. Note that if interactive use is not practical or is ineffective, then these features shall be accessible for cross-development from other platforms. All of the software must be under maintenance for the life of the contract. Updates to the software due to later standards, security requirements, or subsequent development by the Contractor or its sub-contractors shall be included in the contract without additional cost to the Government.

**Table IV**



| *Software Component* | *Features* | *Importance* |
|---|---|---|
| ANSI Fortran95 Compiler | | Required |
| ANSI C Compiler | | Required |
| C++ Compiler | | Required |
| MPI Library | MPI 1.2 | Required |
| | MPI-2 I/O | Required |
| | MPI-2 external interfaces | Desired |
| | MPI-2 Thread safety | Desired |
| OpenMP | | Desired |
| Debugger | Sequential Fortran, C, C++ | Required |
| | MPI Support | Desired |
| | Thread Support | Desired |
| | OpenMP Support | Desired (Required if more than 4 CPUs share the memory space of nodes) |
| Performance Analyzer | Sequential Fortran, C, C++ | Required |
| | MPI Support | Desired |
| | Thread Support | Desired |
| | OpenMP Support | Desired |
| Math Libraries | Standard BLAS, LAPACK, ... | Required |
| | FFT | Desired |

In case the NHRA Computational Platform cannot be partitioned, it shall be required for the offerer to provide a platform for operating system upgrades and application software testing. Even if appropriate NHRA partitioning is available, a test platform may be given additional credit if it permits more effective system utilization.

The NHRA will be installed in the David Skaggs Research Center at 125 Broadway, Boulder, Colorado. The system will be located in the main FSL computer room, Room 2B201. The room has 12" raised floor with removable 2ft x 2ft high-pressure laminate floor tiles. The hallways leading to the computer room have 8" high carpeted floor tiles. Finished floor to drop ceiling height is 8 ft 6 inches. The room has a double door entry with 6 ft total width, 6ft 11 inches total height, and a ramp from the 8" hallway raised floor to the 12" computer room raised floor. The Government requires that the racks composing the NHRA be no taller than 7'2" and support no more than 42U.

The Contractor shall be required to provide at least 50% of the equivalent computational capability of the 768-node Intel cluster available to the current user community during the installation phase of the NHRA. The Contractor can use either the existing hardware or new hardware to ensure that this capability is provided. If the Contractor uses existing hardware, the Contractor must provide maintenance. The Contractor must provide a transition plan as part of the Contractor's proposal. If the Contractor chooses to decommission any part of the existing MSS, the existing data must be available on a read-only basis and the user community must be able to save data to the new system during the transition.

Figure 1 below shows a layout of room 2B201 with the rack space available for the NHRA in green and orange. The hardware in the racks colored in pink and yellow must be maintained by the Contractor(s). The amount of space available is approximately 1100 square feet.

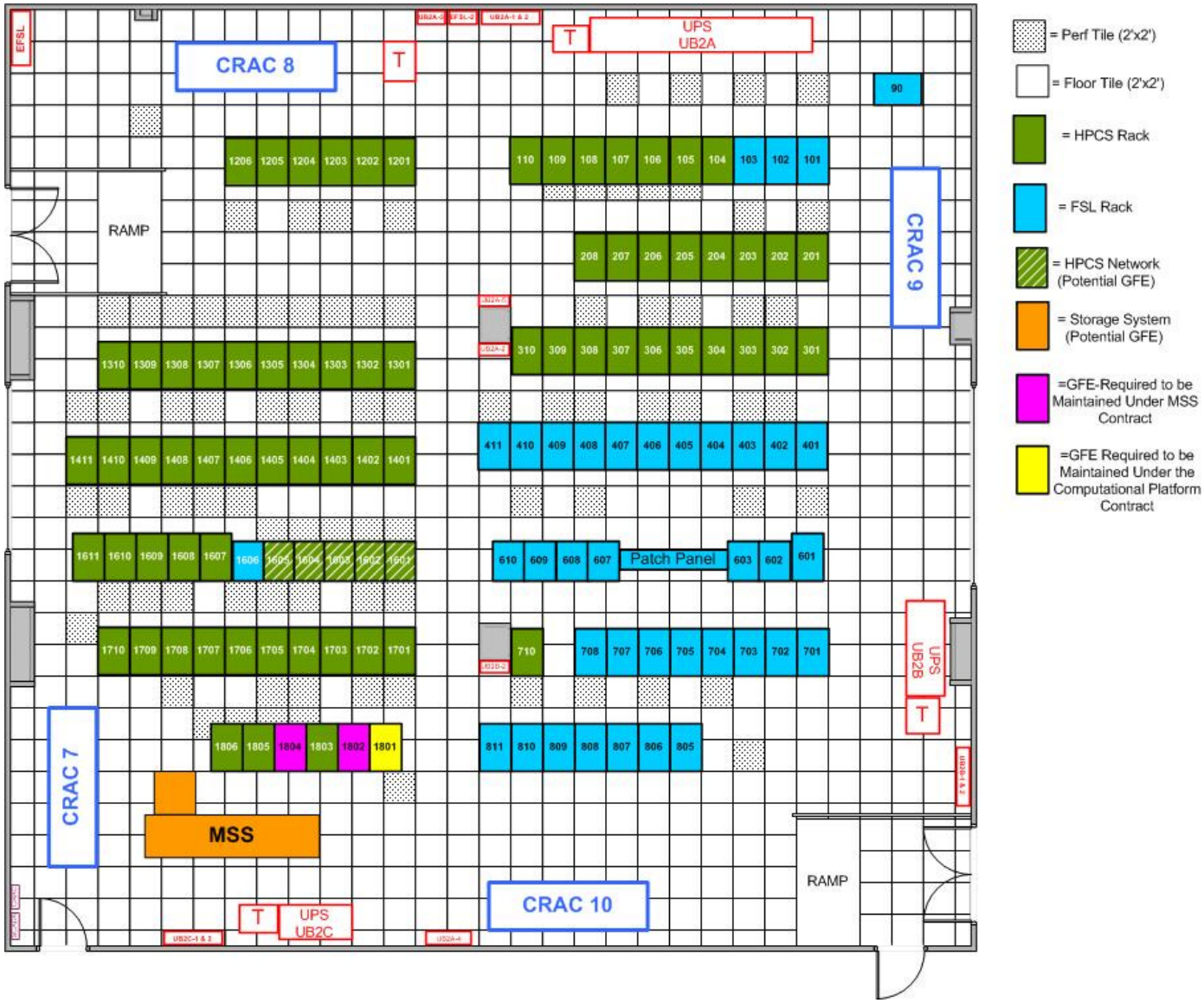


**Figure 1: Room 2B201 Floor Plan**

The amount of power available for the NHRA in 2B201 is 225kVA with commensurate cooling. The power is conditioned with an uninterruptible power supply and backed-up with a motor-generator. Table IV below shows the power circuits that are available for use by the NHRA in 2B201. Changes to the power circuits will require that the Government use funds allocated to this contract. The costs for adding or changing circuits will be deducted from the funds devoted to the contract in the month the Government must make the changes. The costs associated with circuit changes are also shown in Table V. Aside from the power circuit changes, the Government does not intend to make any other changes to room 2B201 or the physical facility.

**Table V**

| Outlet Configuration | Quantity | Action (Change To:) | Cost |
|---|---|---|---|
| L6-20R, 208V | 96 | Two 5-20R, 110V (Requires New Whip) | $750.00 |
| | | Two 5-20R, 110V (Requires New Whip) | $750.00 |
| L5-20R, 110V | 30 | One 5-20R, 110V | $250.00 |
| | | One L6-20R, 208V (Requires 2 Circuits) | $400.00 |
| 5-20R, 110V | 37 | One L5-20R, 110V | $250.00 |
| | | One L6-20R, 208V (Requires 2 Circuits) | $400.00 |

Room 2B201 characteristics are described below:

**Forecast Systems Laboratory**
**Computer Room Infrastructure Specifications**
**FSL Central Computer Facility (2B201)**



| | |
|---|---|
| Dimensions: | 3600 ft² (1100 ft² available to NHRA) |
| | 12" Raised Floor |
| | 8' 6" Ceiling Height (Racks less than 7'2") |
| Flooring: | Tate Access Flooring |
| | 12" Raised Floor |
| Weight Tolerance: | **Computer Room** (ConCore SF 1250 Bolted Stringer) |
| | Concentrated Load: 1250 lbs. |
| | Uniform Load: 300 lbs./ft² |
| | Ultimate Load: 3850 lbs. |
| | Rolling Load: 1000 lbs. (10 Passes |
| | **Surrounding Hallway** (ConCore SF 1250 Cornerlock) |
| | Concentrated Load: 1250 lbs. |
| | Uniform Load: 300 lbs./ft² |
| | Ultimate Load: 3750 lbs. |
| | Rolling Load: 1000 lbs. (10 Passes) |
| Access: | Restricted (Pin Number or Proximity Badge) |
| Power: | Electric Utility Supplied by Xcel Energy |
| | Cutler-Hammer Electrical Distribution Equipment |
| | 480Volt, 3 Phase |
| | On-Site Emergency Generator Backup (Cummins) |
| | Transient Voltage Surge Suppressor (TVSS) Protected |
| | Emergency Power Off (EPO) Switch Protected |
| UPS: | 300 kVA Chloride UPS Systems |
| | (225kVA available to NHRA) |
| | 8-Minute Runtime (Full Load) |

| | |
|---|---|
| Cooling Capacity: | 90-Ton Liebert Downdraft (De-rated for altitude) |
| Fire Protection: | FM-200 Fire Suppression System (Tied to EPO)<br>Vesda Fire Detection System (Tied to FM-200)<br>Cerberus Smoke Detection System (GSA Notification)<br>Sprinkler System (155° F trigger point)<br>CO2 Portable Fire Extinguishers (Class B & C Fires) |
| Monitoring: | Sensaphone SCADA 3000 Monitoring System (Tied to EPO) |
| Cleaning: | Semiannual Professional (Above and Below Floor) |